Solution to Exercise 6.6 (Version 1, 26/10/14)

from Statistical Methods in Biology: Design & Analysis of Experiments and Regression (2014) S.J. Welham, S.A. Gezan, S.J. Clark & A. Mead. Chapman & Hall/CRC Press, Boca Raton, Florida. ISBN: 978-1-4398-0878-8

Exercise 6.6 (Data: courtesy R. Webster, Rothamsted Research, & previously Ecole Polytechnique Fédérale de Lausanne)

The concentrations of several trace metals in a region of the Swiss Jura were quantified by a survey of soil samples at 366 sites (Atteia, Dubois & Webster, 1994). The metals measured (in mg/kg) included cadmium (Cd), chromium (Cr), copper (Cu) and zinc (Zn). The full data set was published in Goovaerts (1997). Here we consider a subset of 207 sample points on a square grid with approximately 250 m spacing. The land use at each sample point was classified into one of three categories (1 = forest, 2 = pasture, 3 = meadow). The unit number (*DSample*), spatial location (*x*- and *y*-coordinates in variates *X* and *Y*, respectively) and land use category (factor LandUse) for each sample can be found in file METALS.DAT along with the concentrations of each metal at each location (variates *Cd*, *Cr*, *Cu* and *Zn*). Analyse the concentration of each metal on an appropriate scale to determine if there are differences among the land types. Are there any metals for which you cannot come to a reasonable conclusion? Plot the co-ordinates of the spatial locations, and consider how you might look for spatial dependence in the residuals. Can you implement your idea? Is there any evidence of spatial dependence?

Solution 6.6

This is an unbalanced set of data: there are 38 samples in category 1 (forest), 43 in category 2 (pasture) and 126 in category 3 (meadow). There are three measurements of copper (Cu) and two measurements of zinc (Zn) concentrations missing. As a first step, we analyse the concentration of each metal using a single factor model with symbolic form (using Cadmium concentration as an example):

Response variable: *Cd*
Explanatory component: [1] + LandUse

We then examine residual plots to see if any further action needs to be taken; we will consider each variable in turn.

(a) Cadmium concentrations (variable *Cd*)

A composite display of residual plots based on standardized residuals is shown in Figure S6.6.1. It is clear that the distribution of the residuals is skewed to the right (histogram bottom left) and that the distribution is not normal (normal plot, bottom right) but with a large number of observations and only three groups, it is not easy to see whether there is any variance heterogeneity. In addition, we might expect to see a greater spread in the meadow category simply because it has many more samples and so has greater probability of sampling the tails of the distribution. In this situation, it is helpful to use boxplots within each category to get a more accurate picture of the distribution. This is illustrated in Figure S6.6.2, where we have taken a sample of random normal values with common variance, allocated them to treatment groups in the same proportions (38:43:126) as in our survey and then plotted them against the fitted value for each group. When plotting the individual points, the variation appears larger for the group with the smallest fitted value, which is the group with 126 observations. Using boxplots, it becomes clear that the interquartile range is actually quite stable across the three groups, it is only the length of the tails that changes.

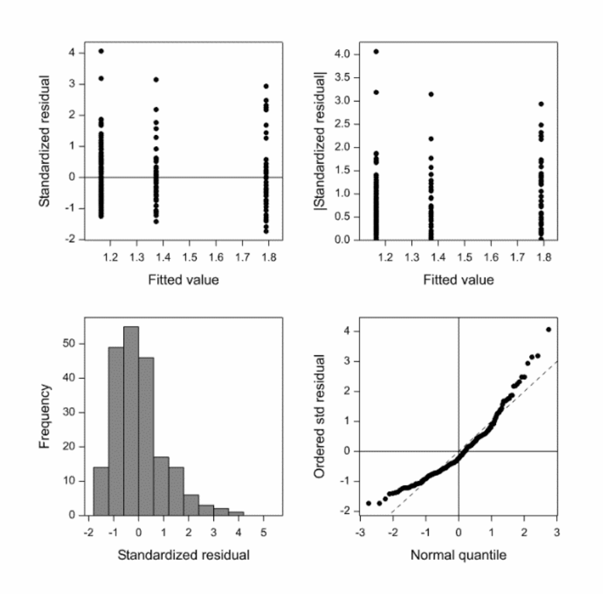

Figure S6.6.1. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate *Cd.*

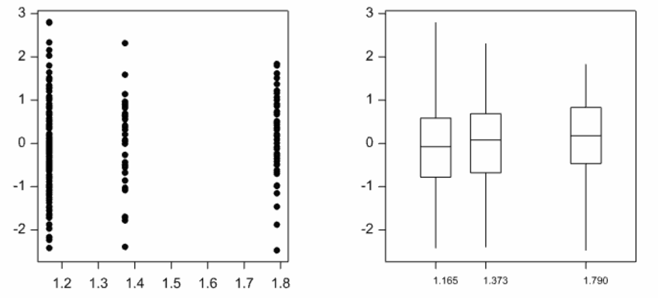

Figure S6.6.2. Random normal samples with common variance allocated to land use categories and plotted against land use mean cadmium values.

A revised set of residual plots, using boxplots to examine distributions within land use categories is shown in Figure S6.6.3. The right skew of the distribution is apparent within each treatment group and the inter-quartile range is larger for the group with the largest mean. We can also calculate the sample mean and variance of residuals within each group, as in Table S6.6.1. The land use group with the largest mean (group 2) also has the largest variance, and Bartlett's test gives strong evidence (P = 0.017) of heterogeneity in variances between groups. We will try to find a transformation such that the residuals conform with the model assumptions.

Table S6.6.1 Table of sample means and unbiased sample variances for three land use categories: Cadmium concentrations.

Land use category	1: forest	2: pasture	3: meadow
Sample mean	1.373	1.790	1.165
Unbiased sample variance	0.6462	1.0900	0.5490

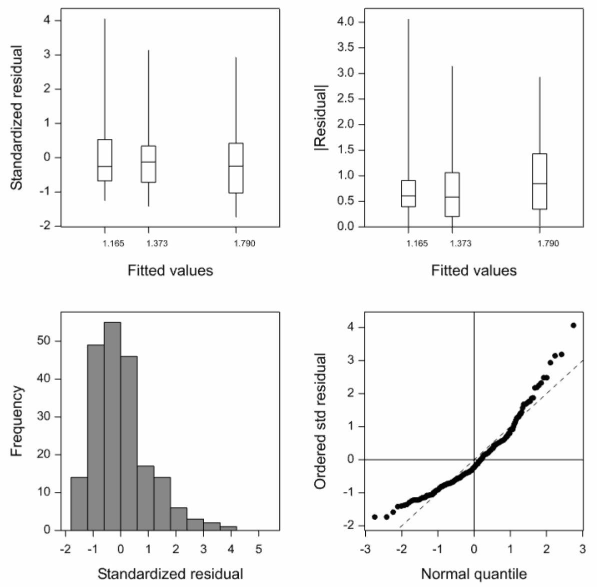

Figure S6.6.3. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate *Cd*, with box-plots to show distributions within treatment groups.

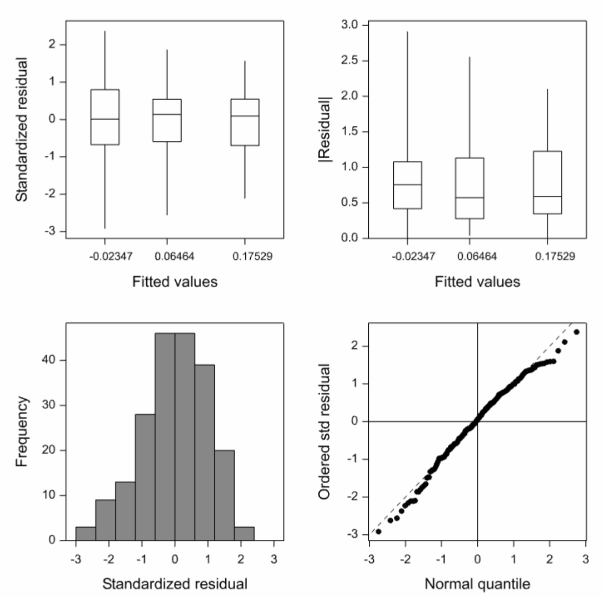

Figure S6.6.4. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate $\log_{10}$(*Cd*).

We first try a log-transformation (here to base 10) and fit the same model to the transformed data. Plots based on the standardized residuals from this analysis are shown in Figure S6.6.4. These plots suggest that the transformation has over-corrected the variance heterogeneity, as the within-group variation is larger for smaller fitted values and the distribution is now skewed slightly to the left. We therefore try a square-root transformation and re-fit the model, leading to the residual plots in Figure S6.6.5. The square-root transformation appears to give a better result. Table S6.6.2 shows the sample means and variances, and a Bartlett’s test shows these values to be consistent with a hypothesis of homogeneity of variances (P = 0.428). The normal plot (bottom right) is closer to a straight line and the distribution of the residuals is more symmetric (although possibly skewed slightly to the right).

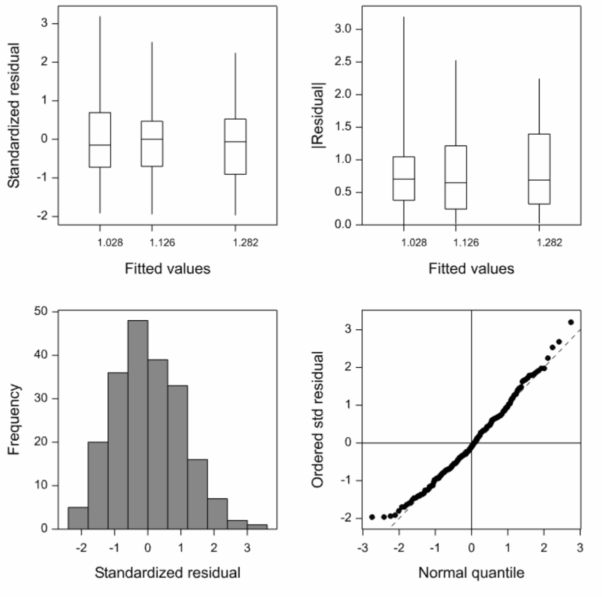

Figure S6.6.5. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate sqrt(*Cd*).

Table S6.6.2 Table of sample means and unbiased sample variances for three land use categories: square-root transformed Cadmium concentrations.

Land use category	1: forest	2: pasture	3: meadow
Sample mean	1.126	1.282	1.028
Unbiased sample variance	0.1093	0.1495	0.1096

These plots seem acceptable, but because of the skew to the right we will try one final transformation, the cube root. Residual plots from this transformation are shown in Figure S6.6.6. This distribution seems more symmetric, the normal plot is closer to a straight line, and the within-group variances are still similar (Table S6.6.3). We will therefore proceed with the analysis on the cube-root scale.

Table S6.6.3 Table of sample means and unbiased sample variances for three land use categories: cube-root transformed Cadmium concentrations.

Land use category	1: forest	2: pasture	3: meadow
Sample mean	1.072	1.168	1.007
Unbiased sample variance	0.0449	0.0567	0.0480

The ANOVA table for analysis on the cube-root scale is Table S6.6.4. There is strong evidence for a difference in population mean cube-root concentration of cadmium among land use categories ($F_{2,204}$ = 8.698, $P < 0.001$). Tables of predicted population means are in Table S6.6.5 with SEDs and back-transformed values. The predicted population mean cube-root concentration of cadmium is greater in pasture than in forest or meadows.

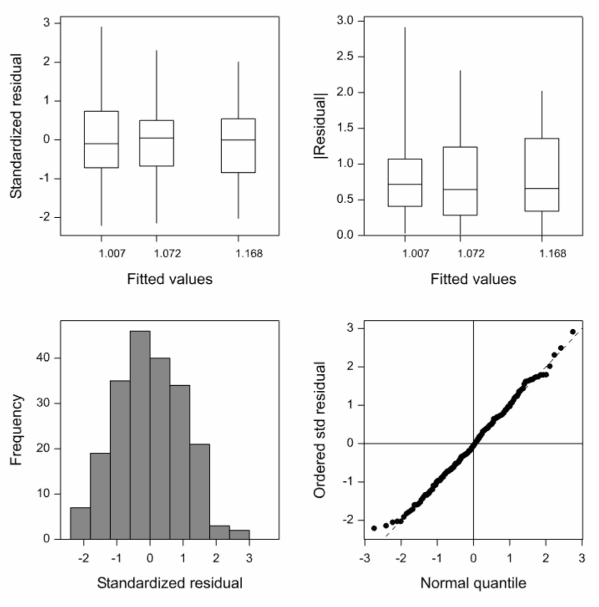

Figure S6.6.6. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate sqrt(*Cd*).

Table S6.6.4 ANOVA table for $Cd^{(1/3)}$ (cube-root transformed concentrations).

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*-value
LandUse	2	0.8561	0.4280	8.698	< 0.001
Residual	204	10.0389	0.0492		
Total	206	10.8950			

Table S6.6.5 Predicted means of cadmium (mg/kg) on cube-root scale for three land use categories with back-transformed value. SEDs (on cube-root scale, 204 df): forest vs pasture = 0.0494; forest vs meadow = 0.0411; pasture vs meadow = 0.0392.

Category	1: forest	2: pasture	3: meadow
Cube-root scale	1.072	1.168	1.007
Back-transformed	1.231	1.595	1.020

(b) Chromium concentrations (variable *Cr*)

We fit the same single factor model to chromium concentrations to obtain the composite display of residual plots (based on standardized residuals) shown in Figure S6.6.7. These plots seem acceptable. The overall distribution is reasonably symmetric, as are the distributions within land use categories. The variances within land use categories are similar (Bartlett's test with $P = 0.807$, see also Table S6.6.6), and the normal plot is very close to a straight line. We will therefore analyse this variable without transformation.

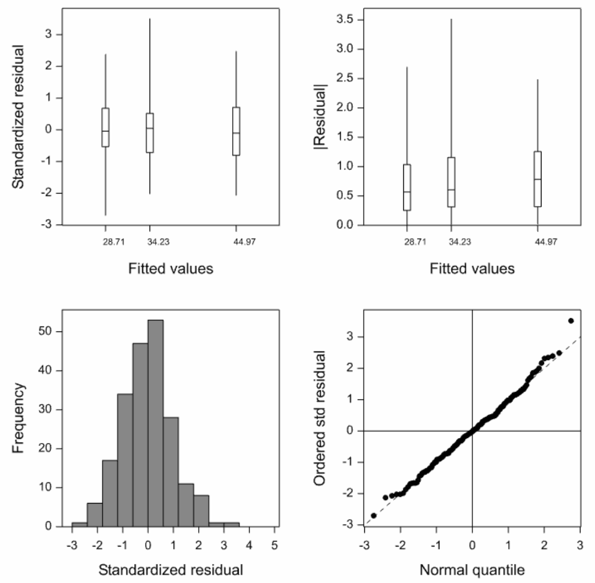

Figure S6.6.7. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate *Cr*.

Table S6.6.6 Table of sample means and unbiased sample variances for three land use categories: Chromium concentrations.

Land use category	1: forest	2: pasture	3: meadow
Sample mean	28.71	44.97	34.23
Unbiased sample variance	91.214	102.057	86.642

Table S6.6.7 ANOVA table for *Cr* (untransformed).

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*-value
LandUse	2	5812.595	2906.298	32.061	< 0.001
Residual	204	18492.175	90.648		
Total	206	24304.770			

Table S6.6.8 Predicted means of chromium (mg/kg) for three land use categories. SEDs (204 df): forest vs pasture = 2.120; forest vs meadow = 1.762; pasture vs meadow = 1.682.

Category	1: forest	2: pasture	3: meadow
Cube-root scale	28.71	44.97	34.23

The ANOVA table is Table S6.6.7. There is strong evidence for a difference in population mean concentration of chromium among land use categories ($F_{2,204}$ = 32.061, *P* < 0.001). Tables of predicted population means are in Table S6.6.8 with SEDs. The predicted population mean concentration of chromium is greater in pasture than in meadows, which in turn is greater than in forests.

(c) Copper concentrations (variable *Cu*)

There are three missing measurements of copper concentrations, so we remove these observations from the data set before analysis. We fit the single factor model to obtain the composite display of residual plots (based on standardized residuals) shown in Figure S6.6.8. The distribution of the residuals is strongly skewed to the right, and the normal plot shows marked curvature. The boxplots of the distributions within each treatment group show that the within-group distributions are skewed and the variance is larger for the groups with larger fitted values. In this case, this pattern is also clear from plot of the individual residuals (not shown). These data clearly do not obey the assumptions underlying the single factor model and so we seek a transformation to improve matters.

We first try a log transformation (to base 10). This improves matters somewhat (Figure S6.6.9): the right skew in the distribution is much smaller and the normal plot is close to a straight line. The within-group variances are more even, although the variance for group 2 (second highest mean on log-scale) appears larger, and this is verified by Bartlett's test (P < 0.001). However, transformation to other scales (we tried square root and reciprocal) does not resolve this problem. We choose to proceed with the analysis on the $\log_{10}$-scale but to treat out results with some caution.

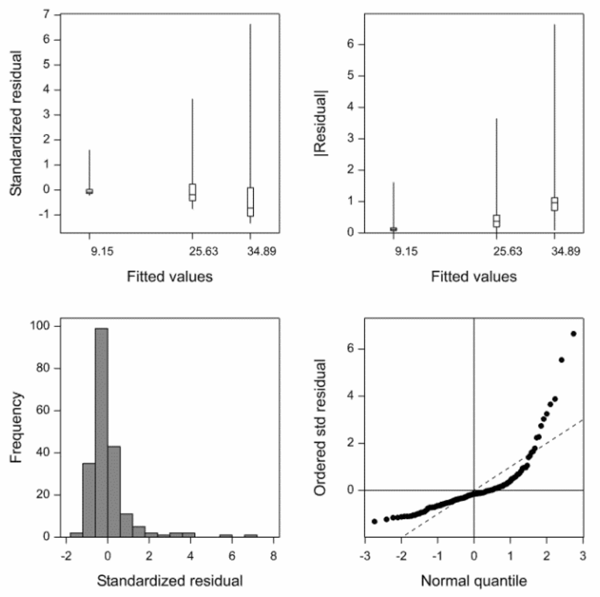

Figure S6.6.8. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate Cu.

Table S6.6.9 ANOVA table for $\log_{10}$(*Cu*) ($\log_{10}$ transformation).

Source of variation	df	Sum of squares	Mean square	Variance ratio	P-value
LandUse	2	5.8459	2.9230	38.232	< 0.001
Residual	201	15.3670	0.0765		
Total	203	21.2129			

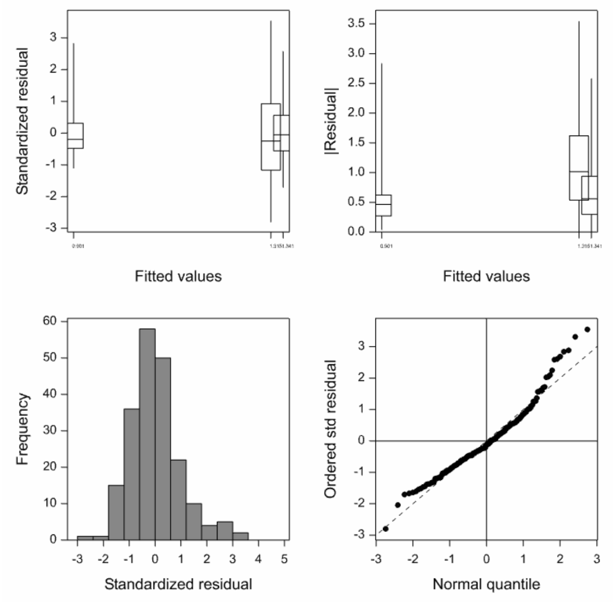

Figure S6.6.9. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate $\log_{10}$(Cu) ($\log_{10}$-transformation).

The ANOVA table for analysis of $\log_{10}$-transformed copper concentrations is Table S6.6.9. This shows strong evidence of a difference between population mean $\log_{10}$-transformed copper concentrations between land use categories. The predicted population means are shown in Table S6.6.10. The predicted population mean $\log_{10}$ copper concentration in forests is smaller than that in pasture or meadows.

Table S6.6.10 Predicted means of $\log_{10}$-transformed copper concentrations (mg/kg) for three land use categories with back-transformed value. SEDs (on $\log_{10}$ scale, 201 df): forest vs pasture = 0.0619; forest vs meadow = 0.0513; pasture vs meadow = 0.0494.

Category	1: forest	2: pasture	3: meadow
$\log_{10}$ scale	0.901	1.315	1.341
Back-transformed	7.95	20.67	21.93

(d) Analysis of zinc concentrations (variate *Zn*)

There are two missing measurements of zinc concentrations, so we remove these observations from the data set before analysis. We again fit the single factor model to obtain the composite display of residual plots (based on standardized residuals) shown in Figure S6.6.10. The distribution of the residuals is strongly skewed to the right, and the normal plot shows marked curvature. The boxplots of the distributions within each treatment group show that the within-group distributions are skewed and the variance is larger for the groups with larger fitted values. In this case, this pattern is also clear from plot of the individual residuals (not shown). These data clearly do not obey the assumptions underlying the single factor model and so we seek a transformation to improve matters.

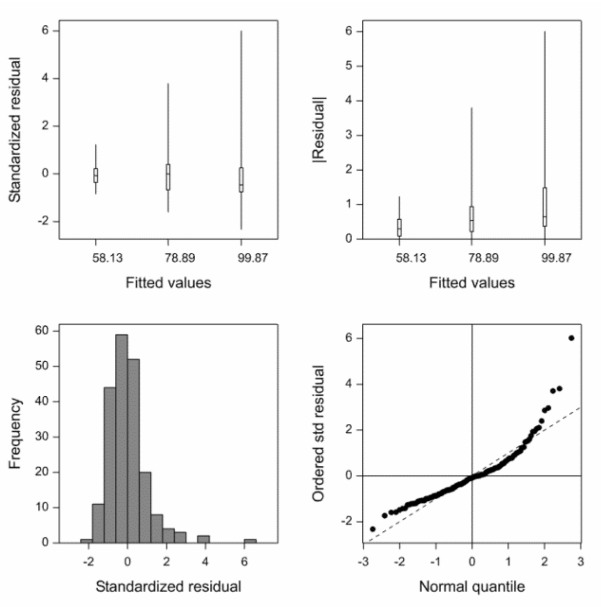

Figure S6.6.10. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate *Zn*.

We first try a $\log_{10}$-transformation, and the resulting residual plots are shown in Figure S6.6.11. The distribution has become more symmetric and the normal plot is close to a straight line. However, there is still a suggestion that the variance is greater for groups with larger mean values. Again, we cannot find a transformation to resolve this issue, and so we proceed with analysis on the $\log_{10}$-transformed values with caution.

Table S6.6.11 ANOVA table for $\log_{10}$(*Zn*) ($\log_{10}$ transformation).

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*-value
LandUse	2	0.8847	0.4424	17.461	< 0.001
Residual	202	5.1176	0.0253		
Total	204	6.0024			

Table S6.6.12 Predicted means of $\log_{10}$-transformed zinc concentrations (mg/kg) for three land use categories with back-transformed value. SEDs (on $\log_{10}$ scale, 202 df): forest vs pasture = 0.0356; forest vs meadow = 0.0295; pasture vs meadow = 0.0284.

Category	1: forest	2: pasture	3: meadow
$\log_{10}$ scale	1.747	1.957	1.869
Back-transformed	55.89	90.59	74.01

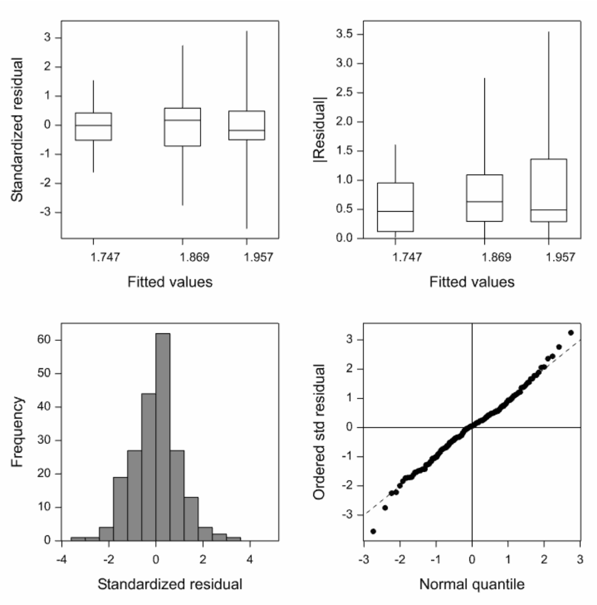

Figure S6.6.11. Composite set of residual plots based on standardized (std) residuals obtained from analysis of variate $\log_{10}(Zn)$.

The ANOVA table for analysis of $\log_{10}$-transformed zinc concentrations is Table S6.6.11. This shows strong evidence of a difference between population mean $\log_{10}$-transformed zinc concentrations between land use categories. The predicted population means are shown in Table S6.6.12. The predicted population mean $\log_{10}$ zinc concentration in forests is smaller than that in meadows, which is smaller than that in pasture.

Overview of analyses

We have found a cube-root transformation for cadmium that makes the residuals appear consistent with normal errors, and chromium appears to be consistent with the model assumptions without a transformation, so the analysis and interpretation for these two variables is straightforward.

For copper concentrations, there appears to be heterogeneity between land use categories that is not related to the population means. It is therefore unlikely that transformation can solve this problem. There might be sub-classes of land use within the more variable group with different means – this would cause extra within-group variation – and if we can identify the sub-classes then we might include them in our analysis. Similarly, there might be one or more additional explanatory variables that should be included in the model to account for within-group differences (see Chapters 14-15). Without any further information, we might use a weighted analysis to analyse this data, but this is beyond the scope of our book. As an ad hoc alternative, as there are only three groups, we could compare groups using t-tests but without assuming a common variance.

For zinc concentrations, the heterogeneity does appear to be related to group means, but is not resolved by the $\log_{10}$ transformation. We might speculate that these data come from a distribution where the variance is related to a power of the mean (such as the negative binomial distribution) and try to identify this distribution and use it in the analysis (like the GLMs in Chapter 18).

Finally, we should remember that with such a large data set, the ANOVA is likely to be able to detect very small differences in population means. It would be sensible to pause to consider whether differences are biologically meaningful before reporting them so that the differences (although statistically significant) can be put into an appropriate context.

Checking for spatial dependence

A plot of the spatial co-ordinates is shown in Figure S6.6.12. The simplest way to check visually for spatial dependence on a grid in two dimensions is a heat map of the residuals, where the colour of the area is related to the value of the residual. Any trend in colour across the plot indicates a spatial trend in the residuals. This type of heat map is shown in Figure S6.6.13 for the residuals from analysis of the cube-root of Cadmium concentrations. There is no obvious evidence of spatial trend here. We could also plot sequences of residuals within rows or columns of the grid to look for serial correlation, but here we choose to plot residuals against their neighbour within rows or columns. These plots are shown in Figure S6.6.14, again for cube-root Cadmium concentrations. This plot shows no evidence of spatial correlation.

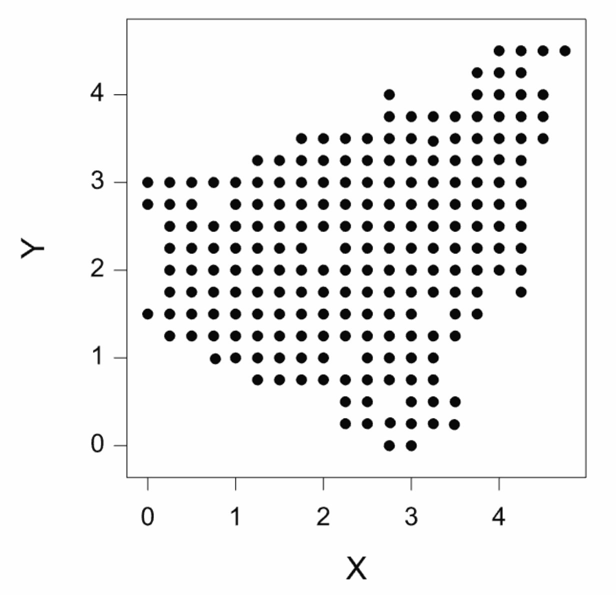

Figure S6.6.12. Plot of the spatial-coordinates at which metal concentrations were measured (actual spacing approximately 250 m).

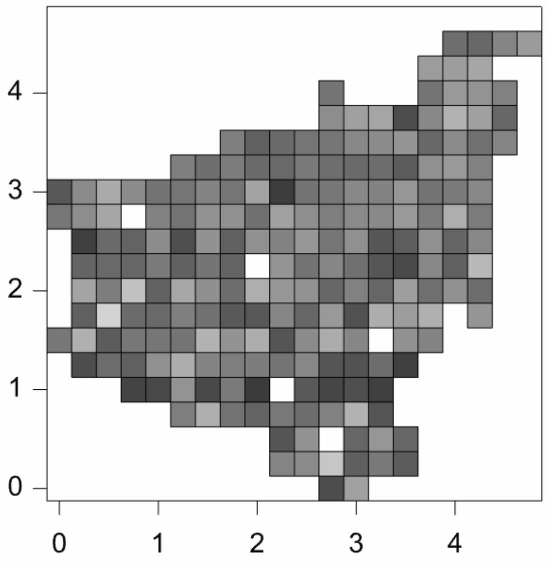

Figure S6.6.13. Heat map of residuals from analysis of cube-root Cadmium concentrations in spatial locations.

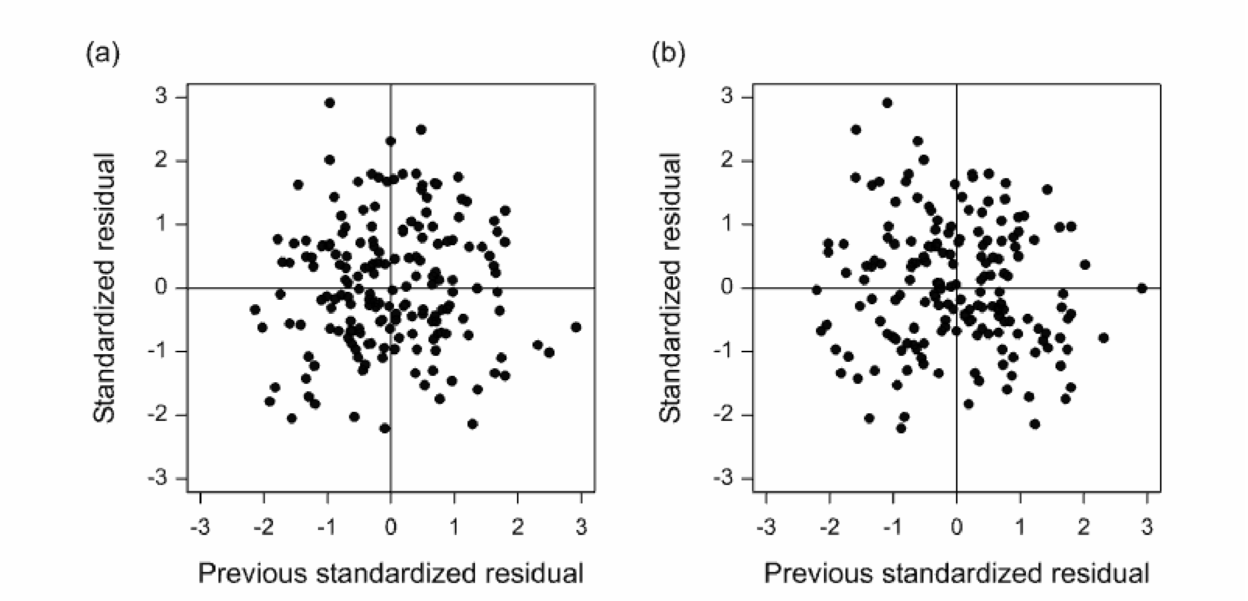

Figure S6.6.14. Plot of residual vs residual from neighbouring (a) row or (b) column for cube-root Cadmium concentrations.

A similar procedure can be followed for each of the other metals. In fact, we do not expect to find any evidence of spatial trend here, as Atteia *et al* (1994) found no evidence of spatial correlation for these metals for samples at 250m distance.